Computer-assisted Learning for Chinese based on Character Families*

John Lee and Chak Yan Yeung

Department of Linguistics and Translation
City University of Hong Kong, Hong Kong SAR
jsylee@cityu.edu.hk, chak.yeung@my.cityu.edu.hk

Abstract. We describe a computer-assisted language learning (CALL) approach for Chinese that is based on character families. This approach exploits word and character embeddings to construct character families that highlight the phonetic regularity and semantic regularity of their member characters. We apply these families in a CALL game where players attempt to combine sub-character components to form characters within a family.

Keywords: Computer-assisted language learning · Chinese · character families · word embeddings

1 Introduction

Chinese language pedagogy is a research area that has been attracting considerable interest [9, 10]. Over 80% of Chinese characters are compound characters, formed by sub-character components called *radicals* [16]. It is therefore advantageous to exploit one's knowledge of the radicals when learning new characters. Indeed, in order to foster awareness of this structural regularity, many computer-assisted language learning (CALL) systems for Chinese offer Scrabble-like games, where players attempt to manipulate radicals to form characters.

As building blocks for characters, some radicals serve as *phonetic radicals* to provide pronunciation cues, and others serve as *semantic radicals* to give hints to the characters' semantic category. Table 1 shows an example. For a systematic review of characters that are phonetically or semantically similar, one can therefore organize characters into "character families" according to phonetic radical or semantic radical [1]. More precisely, members of a "phonetic character family" share a common phonetic radical, and should have identical or similar pronunciation as the radical (Table 2a); members of a "semantic character family" share a common semantic radical, and should be semantically close to the

* This work is partially supported by an Applied Research Grant (#9667175) at City University of Hong Kong, and by the Innovation and Technology Fund (Ref: ITS/389/17) of the Innovation and Technology Commission, the Government of the Hong Kong Special Administrative Region. We thank Lam Xing for implementing the mobile app.

radical (Table 2b). This pedagogical approach is taken, for example, by the *Chinese Radical-Based Character-Family E-Learning Platform.*[1] Empirical studies have shown the benefits brought by knowledge of phonetic and semantic radicals to Chinese language learning [8, 10, 14].

Our goal is to develop a CALL game for Chinese based on phonetic and semantic character families. The game objective is to recognize a character family by examining the radicals that appear in its member characters. To optimize pedagogical effectiveness, it is important to select character families that exhibit a high degree of phonetic or semantic regularity. Not all radicals, however, are reliable in this regard. Only about 38% of the phonetic radicals yield characters with regular pronunciation [18]; only 39% to 58% of the characters are semantically transparent, according to an analysis of primary school materials [2]. As a result, most existing CALL tools had to rely on manual selection. To build a comprehensive game covering a larger number of characters, it would be more efficient to perform automatic selection for character families.

This paper proposes selection methods that exploit word and character embeddings, reports their application in a prototype game, and presents preliminary evaluations. After giving some background on semantic-phonetic compounds and previous work (Section 2), we describe our dataset (Section 3) and game design (Section 4). We then discuss our methods for character family selection (Section 5) and conclude (Section 6).

Compound	**Semantic radical**	**Phonetic radical**
晴	日	青
qíng	*rì*	*qīng*
‘sunny’	‘sun’	‘blue, green’

Table 1. An example semantic-phonetic compound, 晴 *qíng* ‘sunny’, decomposed into its phonetic radical and semantic radical.

(a) Phonetic character family	**(b) Semantic character family**
All members have the phonetic radical 青 *qīng* ‘blue, green’	All members have the semantic radical 日 *rì* ‘sun’
晴 *qíng* ‘sunny’;	晴 *qíng* ‘sunny’;
蜻 *qīng* ‘dragonfly’;	暉 *huī* ‘sunshine’;
請 *qǐng* ‘please’;	曉 *xiǎo* ‘dawn’;
...	...

Table 2. Example phonetic and semantic character families.

[1] Accessed at http://other.allad.com.tw/chinese2/fonts.php

2 Background

After a discussion on semantic-phonetic compounds (Section 2.1), which form the focus of our CALL game, we characterize different degrees of phonetic and semantic similarity (Section 2.2) and review existing games (Section 2.3).

2.1 Semantic-Phonetic Compounds

Among the characters listed in *Shuowen Jiezi*, a Chinese dictionary, 82.5% are semantic-phonetic compounds [7]. It has also been estimated that 81% of the most frequent Chinese characters are semantic-phonetic compounds [11].

A semantic-phonetic compound consists of a semantic radical, which indicates the semantic category of the character; and a phonetic radical, which cues the pronunciation of the character. In the example shown in Table 1, the character 晴 *qíng* 'sunny' is made up of two radicals: its left side is the semantic radical 日 *rì* 'sun', which has a closely related meaning; its right side is the phonetic radical 青 *qīng* 'blue, green', whose pronunciation is similar.

2.2 Phonetic and Semantic Regularity

Each Chinese character corresponds to one syllable, which consists of an onset (or *initial*) and a rime (or *final*). Previous studies classified phonetic similarity into four categories [3]: Two characters may have the same phonemes and same tone (e.g., 青 *qīng* and 清 *qīng*); same phonemes but different tones (e.g., *qīng* and 請 *qǐng*); same rime (e.g., *qīng* and 精 *jīng*); same onset, i.e., alliterating (e.g., *qīng* and 恰 *qià*); or no similarity. In a study on the 3,027 most frequent left-right structured characters, about 33% of these compounds are classified as "same phonemes and same tone" [6].

The spectrum of semantic similarity has also been analyzed, for example as six levels of semantic transparency [2]. A character is considered transparent when its meaning is the same as or directly related to the semantic radical (e.g., 櫃 'wardrobe' and its radical 木 'wood'), or when it belongs to the category of the semantic radical (e.g., 姐 'elder sister' and its radical 女 'female'). A character is semi-transparent when its meaning is only indirectly or loosely related to the radical (e.g., 煙 'smoke' and its radical 火 'fire').

2.3 CALL games for Chinese

The objective in most CALL games for Chinese characters is to form characters using a set of radicals [8]. The board game *Zhōngwén pīnzì yóuxì* [2], for example, offers tiles of radicals in left-right and top-bottom structures that can be combined to form over 2,200 characters. Since these games are designed to review structural regularity, the selection of radicals is primarily based on their productivity and frequency, rather than their ability to highlight phonetic or

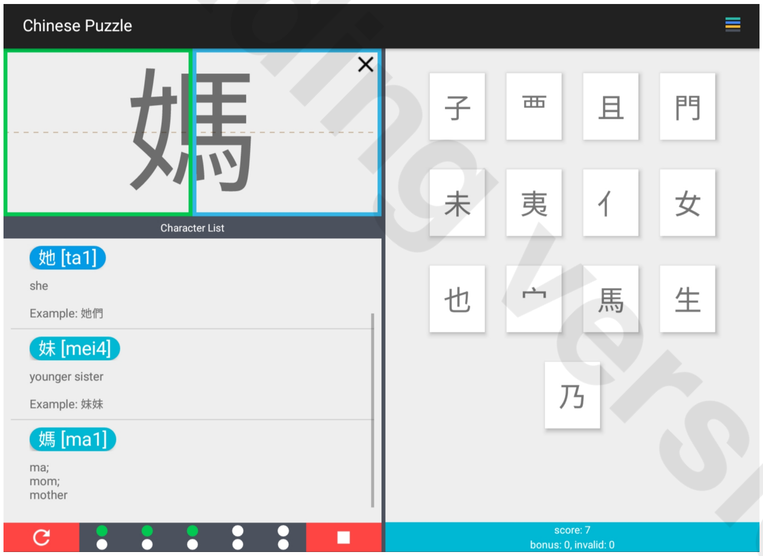
[2] 中文拼字遊戲, published by Sun Ya Publications (HK) Ltd.

semantic patterns among characters. While our proposed approach features a similar game objective, it is distinguished in optimizing the choice of radicals to yield characters that illustrate phonetic and semantic regularities.

3 Data

We constructed a pool of 4,214 characters from the vocabulary lists of the *Hanyu Shuiping Kaoshi* [5] and *Test of Chinese as a Foreign Language* [15], two popular schemes for Chinese pedagogy; as well as from characters that appear in the 40,000 most frequent words in Chinese Wikipedia. We used traditional characters and decomposed them into radicals with HanziJS[3] with the decomposition types "left-right", "top-bottom" and "inside-outside". Of the 4,214 characters, 3,535 were decomposed and they formed the basis of the character families.

We identified the two radicals of each character, and then assigned the character to their respective families. This procedure yielded 616 candidate character families[4], each with an average of 11.1 member characters.



她 [ta1]
she
Example: 她們

妹 [mei4]
younger sister
Example: 妹妹

媽 [ma1]
ma;
mom;
mother

score: 7
bonus: 0, invalid: 0

Fig. 1. Two radicals, 女 *nü* 'female' and 馬 *mǎ* 'horse', have been dragged from the Radical Panel (right) to the Character Box (top left) to form the character 媽 *mā* 'mother'. The Character List (bottom left) displays the valid characters formed so far.

[3] https://github.com/nieldlr/hanzi

[4] We included three additional families taken from the appendix of [12].

4 Game design

Adopting the context-sensitive, just-in-time paradigm for language learning [4], the app randomly chooses a character that appears in the name of the player's current location, as detected by the Google Map API.[5] It then selects either the phonetic or semantic character family (Table 2) of the character.

The member characters of the selected family serve as the "answer set" for the current round of game. The "Radical Panel" displays tiles for all radicals that appear in the answer set, and for a few other radicals that serve as distractors (Figure 1). The player may long-press on any tile to see the English gloss of the radical.[6] Similar to the design of the Upwords variant of Scrabble, the player can drag-and-drop any radical tile to the "Character Box" to stack it on another tile.

The player is to form as many characters and as quickly as possible. When a valid character is formed in the Character Box, it is inserted into the "Character List" (Figure 1) and the player scores. In case of a phonetic family, the score is correlated to the character's category of phonetic regularity (Section 2.2); in case of a semantic family, the score is derived from cosine similarity (Section 5.2). At the end of a round, a summary page shows the total score and the characters formed and missed by the player.

5 Character Family Selection

Among the candidate character families in our dataset (Section 3), we automatically selected those that best illustrate phonetic regularity (Section 5.1) and semantic regularity (Section 5.2).

5.1 Phonetic character families

We compare the pronunciation of each character to the phonetic radical of its family. If the candidate family has at least three member characters with the same phonemes or same rime as the radical, it is accepted for use in the game. Another member character with less phonetic similarity (and hence lower score) is also randomly chosen for inclusion in the answer set.

Evaluation To gauge the difficulty of the game, we conducted a study on 19 students, all native Chinese speakers, at a university in Hong Kong. During a 10-minute period, the subjects played three rounds of game with randomly chosen phonetic character families. The subjects were able to form 50.3% of the characters in the answer set (Table 3). They were more likely to recognize characters with a higher degree of phonetic regularity: 72.2% among characters with the same phonemes or rime, compared to 48.8% among the rest. These results suggest that our game offers a level of difficulty that is sufficient for players to benefit from reviewing the missed characters.

[5] The player can also override the app and directly input the desired character.

[6] The gloss is taken from CC-CEDICT, which was accessed at https://www.mdbg.net/chinese/dictionary?page=cedict.

Characters	% of characters formed
Same phonemes or same rime	72.2%
Other	48.8%
Overall	50.3%

Table 3. Average proportion of characters formed by subjects during games with phonetic character families.

5.2 Semantic character families

We estimate the semantic similarity between a character and its family. Compared to one-hot representations, distributed word representation — representing a word as a vector in a continuous vector space — has been shown to better encode semantic information [13]. We hence derived the character vector and the “family vector”, the representation of the character family, by training word, character and component embeddings with a state-of-the-art algorithm [17].

We defined the family vector to be the average of the embeddings of all member characters, since the component embeddings in [17] did not cover all semantic radicals. As for the character vector, we averaged the embeddings of the 5 most frequent words that contain the character. Compared to the character embeddings already computed in [17], this average led to better performance, likely because it captures the dominant meaning of polysemous characters.

We then computed the cosine similarity score between the character vector and the family vector. The family is accepted for use in the game if it has at least three member characters whose score exceeds 0.5. A character with a lower score is also randomly chosen for inclusion in the answer set.

Evaluation We used the 214 standard semantic radicals in Chinese dictionaries as the gold set of semantic character families. For each candidate character family (Section 3), we computed the average cosine similarity score of its top 10 member characters. The family is accepted if the average score is above a minimum threshold. A threshold of 0.5 produced the best results, at 0.31 precision and 0.95 recall, suggesting that the embeddings trained with [17] are capable of recognizing families with high semantic regularity. Informal analysis on the false positives showed that some of these families also contain member characters with high semantic transparency. This method can thus potentially facilitate the expansion of character learning materials beyond the standard radicals.

6 Conclusion

We have presented a computer-assisted language learning (CALL) approach for Chinese that is based on character families. We described automatic methods to select families whose members are phonetically or semantically similar, and reported preliminary evaluation results. To the best of our knowledge, this is the first effort to design and build a CALL game that highlights both phonetic regularity and semantic regularity among Chinese characters.

References

1. Chen, H.C., Chang, L.Y., Chang, K.E., Chiou, Y.S., Sung, Y.T.: Chinese Orthography Database and Its Application in Teaching Chinese Characters (in Chinese). Bulletin of Educational Psychology (Special Issue on Reading) **43**, 269–290 (2011)
2. Chung, F.H.K., Leung, M.T.: Data analysis of Chinese characters in primary school corpora of Hong Kong and mainland China: preliminary theoretical interpretations. Clinical Linguistics and Phonetics **22**(4-5), 379–389 (2008)
3. DeFrancis, J.: Visible Speech: The Diverse Oneness of Writing Systems. University of Hawaii Press, Honolulu, HI (1989)
4. Edge, D., Searle, E., Chiu, K., Zhao, J., Landay, J.A.: MicroMandarin: Mobile Language Learning in Context. In: Proc. SIGCHI Conference on Human Factors in Computing Systems (2011)
5. Hanban: International Curriculum for Chinese Language and Education. Beijing Language and Culture University Press, Beijing, China (2014)
6. Hsiao, J.H., Shillcock, R.: Analysis of a Chinese Phonetic Compound Database: Implications for Orthographic Processing. Journal of Psycholinguistic Research **35**, 405–426 (2006)
7. Lai, M.D.: Huayuwen Jiaoxue Hanzi Xingshengzi Jiegou Fenxi (in Chinese). The World of Chinese Language **117**, 169–175 (2016)
8. Lam, H.C., Ki, W.W., Law, N., Chung, A.L.S., Ko, P.Y., Ho, A., Pun, S.W.: Designing CALL for Learning Chinese Characters. Journal of Computer Assisted Learning **17**, 115–128 (2001)
9. Lam, H.C.: A Critical Analysis of the Various Ways of Teaching Chinese Characters. Electronic Journal of Foreign Language Teaching **8**(1), 57–70 (2011)
10. Leong, C.K., Tse, S.K., Loh, K.Y., Ki, W.W.: Orthographic Knowledge Important in Comprehending Elementary Chinese Text by Users of Alphasyllabaries. Reading Psychology **32**(3), 237–271 (2011)
11. Li, Y., Kang, J.S.: Analysis of Phonetics of the Ideophonetic Characters in Modern Chinese. In: Chen, Y. (ed.) Information Analysis of Usage of Characters in Modern Chinese (in Chinese). pp. 84–98. Shanghai Education Publisher, Shanghai (1993)
12. Liow, S.J.R., Tng, S.K., Lee, C.L.: Chinese Characters: Semantic and Phonetic Regularity Norms for China, Singapore, and Taiwan. Behavior Research Methods, Instruments, and Computers **31**(1), 155–177 (1999)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proc. International Conference on Learning Representations (ICLR) (2013)
14. Tse, S.K., Marton, F., Ki, W.W., Loh, E.K.Y.: An Integrative Perceptual Approach to Teaching Chinese Characters. Instructional Science **35**, 375–406 (2007)
15. Tseng, W.H.: Huayu baqianci ciliang fenji yanjiu (Classification on Chinese 8000 Vocabulary). Huayu Xuekan **6**, 22–33 (2014)
16. Wang, S.Y.: The Chinese Language. Scientific American **228**, 50–63 (1973)
17. Yu, J., Jian, X., Xin, H., Song, Y.: Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components. In: Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP). p. 286–291 (2017)
18. Zhou, Y.G.: Xiandai hanzihong shengpangde biaoyin gongneng wenti [To what degree are the “phonetics” of present-day Chinese characters still phonetic? Zhongguo Yuwen **146**, 172–177 (1978)